# Segmentation of ancient Arabic documents

Abdel Belaïd, Nazih Ouwayed

# Segmentation of ancient Arabic documents

Abdel Belaïd and Nazih Ouwayed

**Abstract** This chapter addresses the problem of ancient Arabic document segmentation. As ancient documents neither have a real physical structure nor logical one, the segmentation will be limited to textual area or to line extraction in the areas. Although this type of segmentation appears quite simple, its implementation remains a challenging task. This is due to the state of the old document where the image is of low quality, the lines are not straight, sinuous and connected. Given the failure of traditional methods, we proposed a method for line extraction in multi-oriented documents. The method is based on an image meshing that allows it to detect locally and safely the orientations. These orientations are then extended to larger areas. The orientation estimation uses the energy distribution of Cohen's class, more accurate than the projection method. Then, the method exploits the projection peaks to follow the connected components forming text lines. The approach ends with a final separation of connected lines, based on the exploitation of the morphology of terminal letters.

## 1 Introduction

The ancient handwriting is inherently complex because of its irregularity due to the manual aspect of the script. Rarely, the writers used line support (or layer) to write, resulting in sinuous lines of writing. Moreover, because of the calligraphic style of the writing, ligatures are easily introduced between the parts of words and attachment occurred between the words of the successive rows. Adding to this that as the document existed only on paper, updating was done directly on the text itself,

Abdel Belaïd
LORIA, Campus Scientifique, 54500 Vandoeuvre, France - e-mail: abdel.belaid@loria.fr

Nazih Ouwayed
LORIA, Campus Scientifique, 54500 Vandoeuvre, France - e-mail: nazih.ouwayed@loria.fr

which led either to extend the lines in the margins, or adding entire blocks of lines in these margins.

All these artifacts complicate the problem of line segmentation which is essentially contextual in old documents, although most segmentation techniques of modern document are rather ”natural”, seeking essentially parallel alignments of connected components. The problem with this ”contextual” segmentation is new and complex, which is the challenge in research over the last decade.

The literature suggests haphazardly a lot of techniques for extracting lines. Some are more suited than others to the former. We will expose, in the first part of this chapter, a classification of these techniques. Several classification choices are possible, either by the focus type, either by the script type or finally by the method procedure, bottom up or top down. The second section will be devoted to the segmentation of a class of ancient Arabic documents. A further difficulty arises when dealing with Arabic, corresponding mostly to the calligraphic aspect more accentuated in the case of Arabic (see Fig 1 showing different document classes with different kinds of orientation)

Given the failure of traditional methods, we proposed a method for line extraction in multi-oriented documents [45, 44, 46, 47, 48, 49]. The technique has been studied for Arabic documents but can be generalized to any other script that writing is linear. The method starts by an image meshing allowing us to progressively and locally determine the orientations. The orientation is estimated using the energy distribution of Cohen’s class on the projection histogram profile. This local orientation is then enlarged to extract the orientation areas. Afterwards, the text lines are extracted locally in each area basing on the connected components follow-up. Finally, the connected components that touch in adjacent lines are separated..

The chapter is organized as follows. In the section 2, a brief state of the art is given concerning the segmentation approaches. The multi-skew detection algorithm is detailed in section 3. We present some experimental results in section 4 and conclusion and the future trends of this work will be given in section 5.

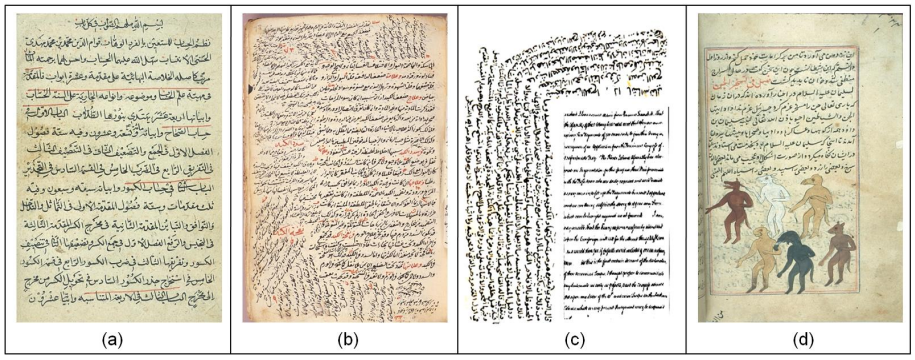


**Fig. 1:** Examples of four categories of handwritten ancient document: (a) Mono-oriented, (b) Multi-oriented, (c) Multi-scripts, and (d) heterogeneous.

## 2 Previous work

The literature mentions a quantity of approaches proposed for document line segmentation. Some of them are top-down while others are bottom-up. .

Top down methods start from the whole image and iteratively subdivide it into smaller blocks to isolate the desired part. They use either an a priori knowledge on the documents such as interline or inter column spaces, or a document model to reach such a segmentation. The localization of white separations is generally done by analyzing the projection histogram profile, either by analyzing vertical stripes like in [1], or by shredding the interline surface with local minima tracers like in [2], or finally by using a vector distance between histogram peaks and pixels, as in [3]. To face the inclination problem, one uses Hough transform that considers the whole image composed of straight lines [4]. Given the failure of these global methods, other researchers are trying to use a knowledge model, as DMOS model proposed by [5] which consists of a grammatical formalism position to model the document structure, or the model of [6] corresponding to a vectorization based algorithm, parametrized by some line features such as angle, length, etc. Nicolas, Paquet and Heutte [7] propose an IA problem solving framework using production systems.

Bottom up methods deal with noise problems and writing variation. Most methods of line extraction in handwritten documents are bottom-up. The connected component based methods are the mainstay of the bottom up approaches. They are clustered into bigger elements such as words, lines and blocks. In each research, simple rules are used in a different way. These rules are based on the geometric relationships between neighboring blocks, such as distance, overlap, and size compatibility. The difference between the different works lies in their capabilities to cope with space variation and influence of the script and the writer peculiarities. Several approaches for clustering connected components have been proposed in the literature, such as K_NN, Hough transform, Smoothing, Repulsive-attractive Network, Minimal Spanning Tree (MST) and deformable models.

Clustering methods related to the notion of mutual neighborhood have been considered in the clustering literature, as in [8] where the clustering is operated on different kinds of textual blocks extracted from vertical strips, or in [9] where a perceptual grouping based on the ”Gestalt theory” principles, such as proximity and similarity, is operated, or finally in [10] where the grouping is based on the text line alignments. Hough transform is also used in bottom up approaches. The main questions are related to the voting points, the most representative of the text lines. In [11], the voting points correspond to the center of gravity of connected components. In Pu and Shi [12], they correspond to the minima of the connected components, located in a vertical strip on the left side of the image. In [13], the voting points correspond to character blocks which size is estimated from the average of the character sizes in the document.

The smoothing technique (Run Length Smoothing or RLS), is to darken the small spaces between the consecutive black pixels in the horizontal direction, which leads to connect them. The boxes which include the successive connected components in the image, form the lines. In [14], a fuzzy run length algorithm is used. In [15],

lines are extracted by applying RLSA, adapted to a gray scale image. Instead of connecting a series of white and black pixels, the gradient of the image is expanded in the horizontal direction with a tilt angle that varies between $\pm 30^{\circ}$.

The repulsive attractive network is a dynamic system to minimize energy, which interacts with the textual image by attractive and repulsive forces defined on the network components and the document image [16]. Experimental results indicate that the network can successfully extract the baselines under significant noise in the presence of overlaps between the ascending and descending parts of characters in two lines.

Considering the connected components in a document as the vertices of a graph, we can obtain a complete undirected graph. A spanning tree of a connected graph is a tree that contains all the vertices of this graph. A minimal spanning tree of a graph is that spanning tree for which the sum of the edges is minimal among all the spanning trees of this graph. A minimal spanning tree of a graph can be generated with Kruskal algorithm. In this algorithm, the tree is built by inserting the remaining unused edge with the smallest cost until all the vertices are connected [17].

The deformable model is an analytical approach which can act interactively on the modeling. It allows to change (in time and space) the model representation of the model towards the solution of the minimization problem introduced in the modeling. Concretely, this leads to introduce a term of time evolution in the minimization criterion, which allows, each time, to influence the prior model when necessary, and to readjust to a better solution. Early work in this area are those of Kass, Witkin and Terzopoulos[19]. In the case of a 2D image, the deformable contour model is used to find an existing object. The process is iterative. From an initial contour, a mechanism of deformable contour is applied to change this form so that it is the target area. The evolution mechanism is an energy function. The target area will be found by minimizing this energy. Several deformable contour models exist in the literature. Here are a few examples: the parametric active contour model (snake [19], the geometric snake [20], the Level Set method [21, 22, 23], the B-spline or B-snake [24] and the Mumford-Shah model [41]).

Table 1 summarizes all the methods mentioned, divided according to 15 criteria: lines types (straight, oriented and cursive), materials types (printed, handwritten, multi-oriented interval orientation (IO), Latin, Chinese, Indian, Arabic, Persian, Urdu, image level (C: Color, G: Gray and B: Binary) and mesh. All these approaches are either too general, proceeding by projection or by alignment search, or too local, operating by connected component following. They find here their limits facing to the poor quality and multi-orientation of documents. Most of these techniques have been applied to documents with a single orientation. The adaptation of these approaches is impossible if we want to extract all directions.

| Class | Category | Authors | Line Type | | | Material Type | | | | | | | | | | | | | Meshing |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Straight | Oriented | Cursive | Printed | Handwritten | Multi-oriented | I. O. | Latin | Chinese | Indian | Arabic | Persian | Urdu | Level | |
| Top-Down | Projection based | [Nicolaouall209] | x | x | | | x | | | x | | | | | | G,B | |
| | | [Bennasri et al. 99] | x | x | | | x | | | | | | x | | | B | x |
| | | [Shapiro et al. 93] | x | x | | | x | | | x | | | | | | B | |
| | | [Antonacopoulos et al. 04] | x | | | x | | | | x | | | | | | B | |
| | Document Model | [Nagy et al. 88] | x | | | x | | | | x | | | | | | B | x |
| | | [Lu et al. 04] | x | x | | x | | | | x | | | | | | B | x |
| Bottom-Up | k-NN | [Likformanfaure1994] | x | x | x | | x | | | x | | | | | | G,B | |
| | | [Zahour et al. 04] | x | x | x | | x | | | | | | x | | | B | |
| | | [Zahour et al. 07] | x | x | x | | x | | | | | | x | | | B | |
| | | [Zahour et al. 08] | x | x | x | | x | | | | | | x | | | B | |
| | | [Feldbach et al. 01] | x | x | | | x | | | x | | | | | | G | |
| | Hough Transform | [Likforman et al. 95] | x | x | | | x | | | x | | | | | | B | |
| | | [Pu ert al. 98] | x | x | x | | x | | | x | | | | | | B | |
| | | [Malleron et al. 09] | x | x | | x | x | | $\pm 45^\circ$ | x | | | | | | B | |
| | | [Louloudis 09] | x | x | x | x | x | | | x | | | | | | G,B | |
| | Smoothing | [proceeding8] | x | x | | | x | | | | | | | | | B | |
| | | [Bourgeois et al. 01] | x | x | | x | x | | | | | | | | | G | |
| | Repulsive-attractive | [Oztop et al. 99] | x | x | x | | x | | | x | | | x | x | | G | |
| | MST | [Yin et al. 08] | x | x | | | x | | | | x | | | | | B | |
| | | [Nicolas et al. 04] | x | x | | | x | | | x | | | | | | B | |
| | Snake | [Bukhari et al. 09] | x | x | x | x | x | x | | x | | | x | | x | G,B | |
| | Level set | [Li et al. 08] | x | x | x | x | x | x | | x | x | x | x | | | B | |
| | Mumford-Shah | [Xiaojun et al. 09] | x | x | x | | x | x | | x | | | | | | G | |

**Table 1:** Line segmentation bibliography

## 3 Overview of the proposed system

Given the failure of traditional methods, we proposed a method for line extraction in multi-oriented documents. The technique has been studied for Arabic documents but can be generalized to any other script that writing is linear. The method is based on an image meshing that allows it to detect locally and safely the orientations. These orientations are then extended to larger areas. The only assumption is that initially the central part of the paper is horizontal. The orientation estimation uses the energy

distribution of Cohen's class, more accurate than the projection method. Then, the method exploits the projection peaks to follow the connected components forming text lines. The approach ends with a final separation of connected lines, based on the exploitation of the morphology of terminal letters.

## *3.1 Image Meshing*

In this step, the document image is partitioned into small meshes. The mesh size is generated, based on the idea that a mesh must approximately contain 3 lines, so as to produce a projection histogram profile that is representative of the writing orientation. To find the lines, the Active Contour Model (or Snake) is applied. The traditional external energy has some limitations such as the edge initialization near the contour and the poor convergence to regions with concavities. For that reason, Xu et al. [30] developed a new kind of external energy that permits the snake to start far from the object, and forces it into boundary concavities. This energy is named gradient vector flow, or GVF.

In our application, the major axis (equal to the first harmonic of the Fourier descriptor) of the connected components is used as the initial Snake. We used the GVF as external energy and a null internal energy. To detect the alignments, some morphological operations such as dilation and erosion are first applied to the initial image (see Figure 2.b) to expand the edges. Then, the major axis of each connected component is determined using the Fourier descriptors [31] (see Figure 2.c). Finally, the energy minimization mechanism is operated on the snake to deform and push it to the text edge, more or less similar to the connected component skeleton (see Figure 2.d). To ensure that the lines will be detected, we increment the size of the major axis by a threshold equal to the quarter of the average width of the connected components. This threshold is obtained by experiments. Finally, the connected components that belong to the same line are grouped to form the lines (see Figure 2.e). Figure 4.b shows the results of the automatic meshing of the document presented in the figure 4.a.

In order to reduce the running speedup, we discard the meshes containing few pixels because their inclination is insignificant. If a mesh contains some text (i.e. few connected components) and thus no noise, it is automatically merged with the neighbor meshes.

## *3.2 Orientation Area Extraction*

### 3.2.1 Orientation Estimation

As the lines are wavy, the orientation is first searched in small meshes where it is more likely to have fragments of straight lines. Traditionally, the projection his-

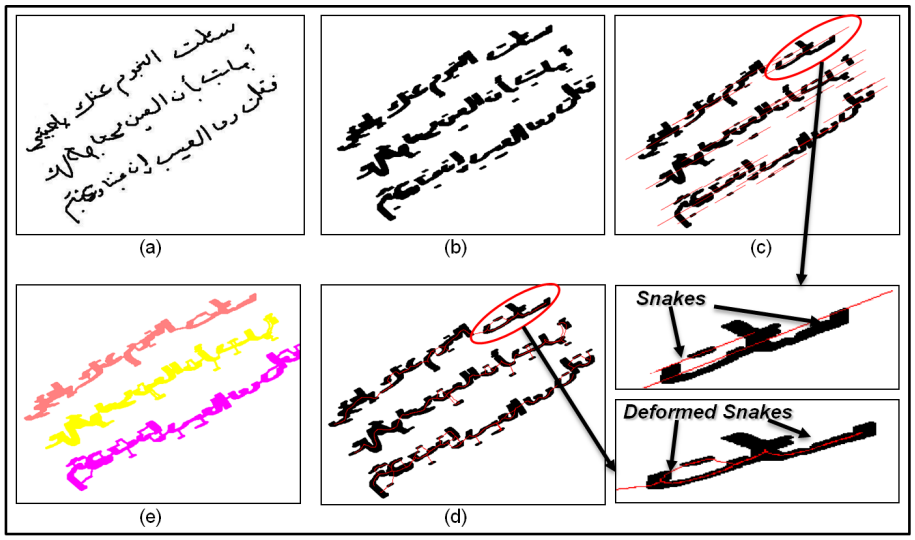


**Fig. 2:** Application of the Snake for line detection, (a) initial image, (b) dilation and erosion of the image, (c) major axis drawn for each connected component of the lines. The ellipse encapsulates the initial connected component of (b). (d) shows the distorted snake of (c). (e) gives the final result showing the connected components gathered in each line.

togram profile is employed along different orientation angles to determine the local orientation by the calculation of the difference between peaks and valleys. However, we observed that technique fails for Arabic which individual word parts (PAWs) can be oblique while the global word is horizontal. To face this problem, we have examined other features to better analyze the histogram function. We then used the energetic time-frequency distributions on the histogram considered as a signal.

### 3.2.2 Time-Frequency Distributions

To have a more robust estimator, we therefore considered using a time-frequency representation of the histogram projection. This distribution best responds that projection to the peaks generated by the lines, translating their presence in high energy. It is becoming less sensitive to false maxima disrupting calculating depths of the peaks in the histogram which his orientation significant. We used the Cohen's class distributions which are quadratic and verify the invariance property by temporal or frequency translation. Each member of this class is distinguished by a kernel which has a determinant role in the quality of the provided images and in the properties it verifies. We limited to the Wigner-Ville distribution (WVD) which their properties allow to be more reactive to the presence of histogram peaks than the others distributions of the Cohen class [44].

Traditional approaches of signal processing such as Fourier transform can not study the signal variation over time and frequency. The energetic time-frequency distributions go beyond what these approaches allow by analyzing the non-stationarity of a signal and distribute the energy of a signal in time and frequency.

According to [37], the energy $E_x$ of a signal $x(t)$ is defined as

$$E_x = \int_{-\infty}^{+\infty} |x(t)|^2 \, dt = \int_{-\infty}^{+\infty} |\widehat{x}(f)|^2 \, df \tag{1}$$

where $\widehat{x}(f)$ is the Fourier transform of the signal $x(t)$. The value $E_x$ is quadratic. For this reason, the time-frequency distributions must keep this property.

**Cohen's Class:** in 1966, Cohen [33, 34, 35] proved that a significant number of time-frequency distributions can be seen as particular cases of the following general expression:

$$C_x(t,f) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \phi_{dD}(\tau,\xi) A_x(\tau,\xi) e^{j2\pi(t\xi - f\tau)} d\xi d\tau, \tag{2}$$

where $A_x(\tau,\xi)$ is the ambiguity function defined by:

$$A_x(\tau,\xi) = \int_{-\infty}^{+\infty} x(t+\tau/2)\, x^*(t-\tau/2)\, e^{-j2\pi\xi t} \, dt$$

The Cohen's class contains all the time-frequency distributions that are covariant under time- and frequency-shifts. The members of this class are identified by a particular kernel $\phi_{dD}(\tau,\xi)$, (expressed here in the delay-Doppler plane $dD$) which determines their theoretical properties [36, 38, 37] and their practical readability.

We want to use these distributions on the signal representing the histogram projection profile in each mesh, in order to estimate its orientation. The Cohen's class distributions are used to estimate the orientation because when computing the projection histogram of a document along one direction of projection, we obtain, if this direction is the real orientation of the document, a histogram in which each line leads to a clearly localized local maximum. Each block of the document leads in the projection histogram to a succession of periodic peaks and valleys, whose period is relatively constant. This periodic succession is delimited by the block size ("time" support) and oscillates at a frequency determined by the space width between the lines. As all the pixels are accumulated in the same positions, the local maxima have higher energy levels than with other projection directions. This explains why we can estimate the orientation of a document by seeking the projection angle for which the time-frequency distribution localizes a large energy level on a small area of the time-frequency plane. For example, Figure 3 shows the increase of the maximum of the Wigner-Ville distribution when the number of peaks and valleys increases and when the valleys become wider.

To estimate the orientation angle, we use the analytic signal $x_a(t)$ of the centered squared root of the projection histogram $x(t)$ of the document. The analytic signal is the signal $x(t)$ without its negative frequencies. The histogram $x(t)$ is determined

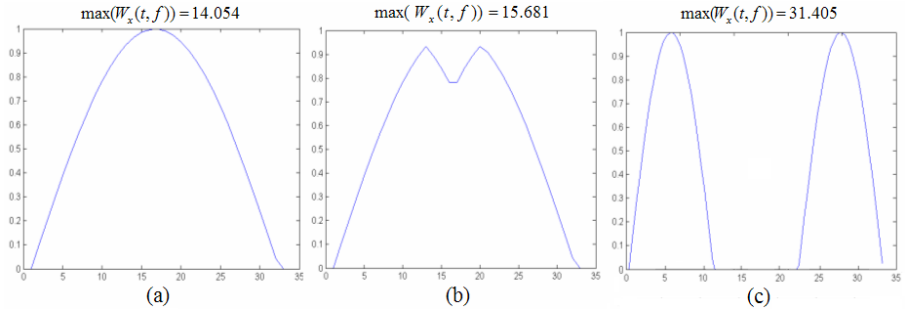


**Fig. 3:** Examples of maximum value of the Wigner-Ville distribution obtained for different projection histogram profiles when the number of peaks and valleys increases and when the valleys become wider.

by projecting each document with a chosen orientation. To calculate all possible projection histograms, we turn the image around its center of gravity (which gives us a point deduced from the image content and not from its size and framing) and we choose the horizontal axis as an arbitrary reference for the zero degree angle. Then, we compute a time-frequency representation for the squared root of each projection histogram, whose average has been removed. The angle corresponding to the histogram with the highest maximum value of its time-frequency representation is chosen as the estimated angle of the document.

Kavallieratou et al. [42, 43] already used the Wigner-Ville distribution (DWV) to estimate the overall direction of Latin documents printed or handwritten. In this work, we wanted to first determine the properties of time-frequency representations that seem desirable for such an application, to then establish a list of performances, the performances were finally evaluated.

### 3.2.3 Orientation Area Expanding

To extend the areas of orientation, we examine the orientations in neighboring meshes and proceed to an extension or a correction. Considering the writing direction in Arabic, we examine pairs of neighbors along three right-left directions: straight, sloping upward and sloping down. The two neighbor meshes are merged if the orientation of the global mesh is equal to one of them, otherwise the orientations are maintained in both meshs. The operation is repeated for all the document meshes. After this step, the zones will be constructed.

When a mesh contains several orientations, the mesh orientation will be erroneous. To detect this phenomenon, we observe the orientation of the horizontal (resp. vertical) surrounding meshes which have different angles. Since this case arises inside the main horizontal (resp. vertical) writing, the vertical (resp. horizontal) projection profile is used to resolve this case. The first minimum value in

the projection profile is looked for from the right representing the end of the first inclination ($I_m$ minimum index). Then the mesh is divided at $I_m$ into two meshes.

Being applied automatically, the initial paving edges can cross the connected components creating problems (false maxima) in text line detection. The incorrect paving exists only in the horizontal and the vertical zones. We need to correct the position of these edges by proceeding a horizontal or vertical shift in order that the local paving covers the local connected components. In the horizontal (resp. vertical) area, the edge that divides two consecutive rows (resp. columns) is moved to the nearest position in these rows (resp. columns) when the horizontal (resp. vertical) projection vector for each of their two consecutive meshes has a minimum value (see Figure 4.g).

### *3.3 Text Line Extraction*

The text line follow-up starts in the first window on the right side of the page. The algorithm starts by looking for the new maxima (see Figure 5.a). Each peak represents the starting point $P_s$ of the orientation line $bl_j$. The ending point $P_e$ of the orientation line is calculated using the $P_s$, the orientation, the width and the height of each window (see Figure 5.b). The orientation line $bl_j$ is calculated basing on the two points $(P_s, P_e)$ and the orientation of the window. The connected components that belong to a baseline are looked for construct the text line (see Figure 5.c).

A step of text line correction follows the text line detection to assign the non-detected components and the diacritical symbols to the appropriate text line (see Figure 5.c and d). A distance method is used to address this problem. First, the distance between the centroid of non-detected component or diacritical symbol $C_i$ and the text line is calculated. $C_i$ is assigned to the text line $l_j$ if $d_{ci,lj} < d_{ci,lj+1}$ else to $l_j + 1$.

### *3.4 Connected Line Separation*

The connections occur between two successive lines when their characters touch. Often, these connections are made between ascenders in the lower line and descenders in the higher line. Table 2 lists the four categories of connection in Arabic: a) a descender with right loop, connects a vertical ascender, b) a left descender with a loop, touches a vertical ascender, c) a right descender touches the higher part of the loop of a character, and d) a left descender connects the higher part of the lower curve of a letter.

In all connection cases, we note the presence of a descender connecting a lower end letter. The descenders are grouped into two categories: (a,c) where the descender of the line starts from the right, and (b, d) where the descender of the line starts from

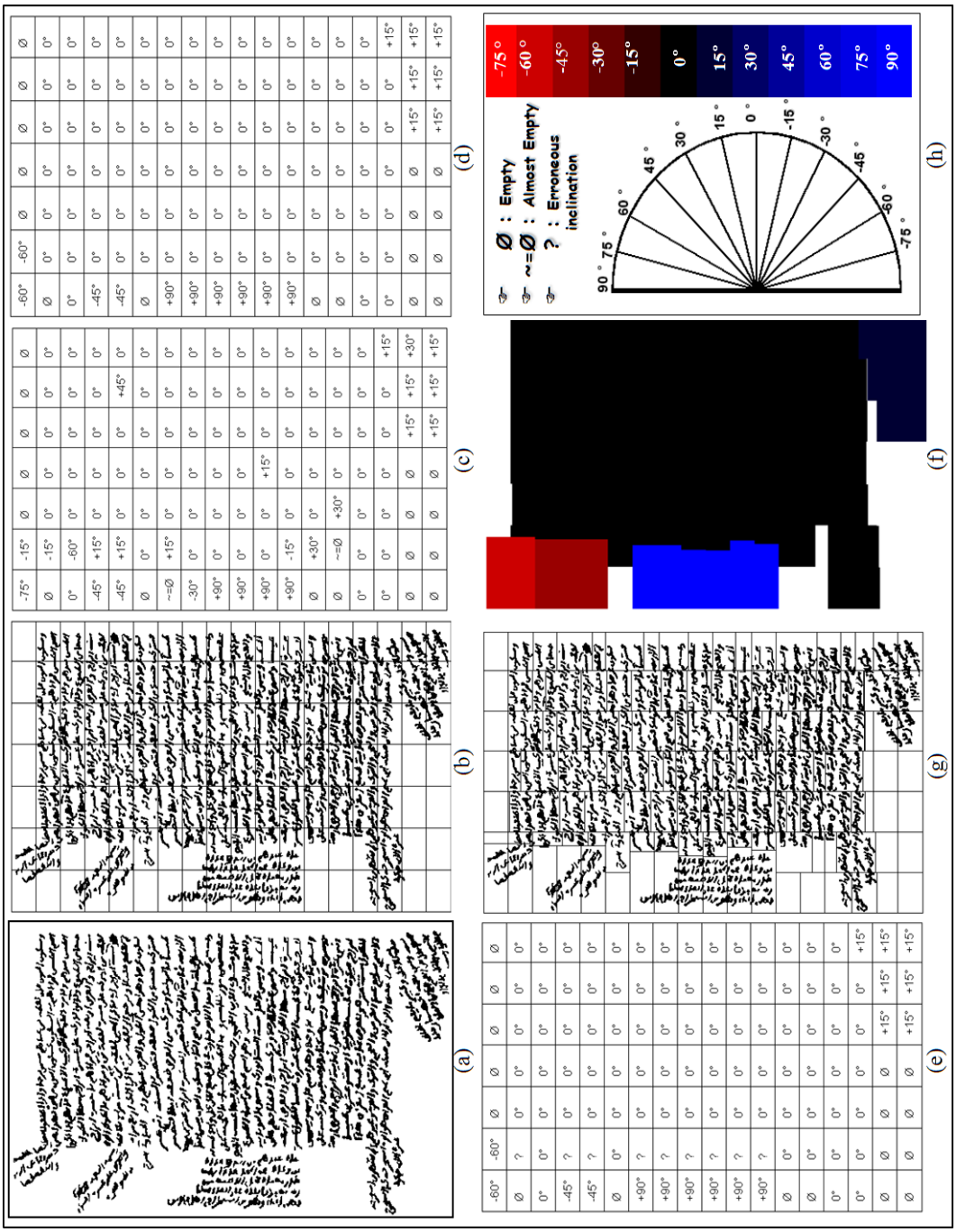


**Fig. 4:** The results for the different steps of the multi-skew detection approach.

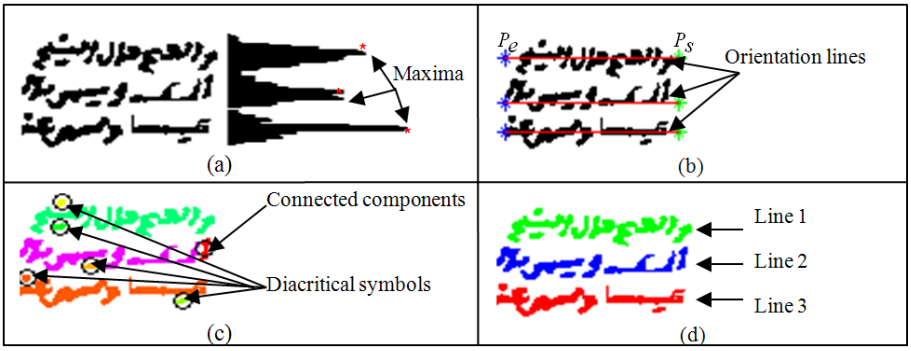


**Fig. 5:** Text line detection steps for a window, (a) maxima detection, (b) orientation lines estimation, (c) assignment of each connected component and diacritical symbol to its appropriate line, (d) extracted lines.

| Type | Area of connexion | Example |
|---|---|---|
| a | | |
| b | | |
| c | | |
| d | | |

**Table 2:** Four types of connexion observed in Arabic handwritten documents.

the left. To streamline the work, the analysis focuses on the connection areas. (see Figure 6 ).

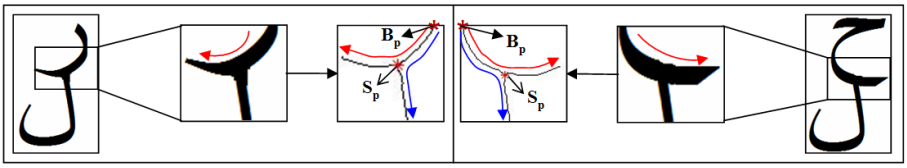


**Fig. 6:** Connection areas and direction of descenders (right direction indicated by the red arrow, and the erroneous direction, by the blue arrow).

The method starts by extracting in the two lines, the connected component created by the connection between the two successive lines (see Figure 7.a). Then, the intersection points of each connected component is detected (see Figure 7.b, the points are in red). An intersection point is a pixel that has at least three neighboring pixels. As in the case chosen, the connection occurs at a single point of intersection

$S_p$ close the minimum axis (valley between two lines, see Figure 7.c). Thus, the point $S_p$ is the nearest point of the minimum axis (see Figure 7.d). We then look for the starting point of the ligature, $Bp$, which is generally the highest point, near the baseline of the top line. Then, from this point, the method is to follow the descending character (i.e. its skeleton, see Figure 7.f). The following continues beyond the intersection point respecting an angular variation corresponding to the curvature of the descending character.

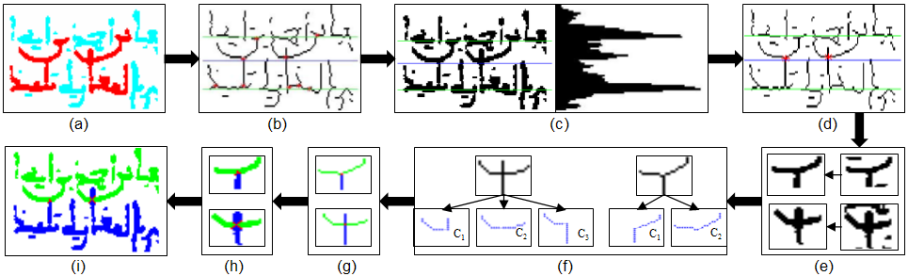


**Fig. 7:** Different steps of the separation of connected components.

Due to the symmetry of the curve branches, the value of the orientation angle must always be positive. For example, in Figure 8, the angular variances are Var $(C_{1+2}) = 703,19$, $Var(C_{1+3}) = 299$, $Var(C_{1+4}) = 572,37$. In this example, the minimum angular variance $Var(C_{1+3})$ is given by the correct direction to follow. Figure 9 illustrates the effectiveness of the algorithm on a representative sample of 12 arbitrarily chosen connected components from 640 occurrences found in 100 documents.

## 4 Experiments and Discussion

To study the effectiveness of our approach, we have experimented on 100 Arabic ancient documents containing 2,500 lines. These documents belong to a database stemmed from web sites of the Tunisian National Library, National Library of Medicine in the USA and National Library and Archives of Egypt. The tests were prepared after a manual areas and lines labeling step of each document. The rotation angle examined during these experiments ranged from $-75°$ to $+90°$. The execution time is measured from the meshing phase until the line separation phase. It depends on the document and the mesh sizes. The tests were performed on a PC with a Pentium M 1.4 GHz and a cache of 1 GB in Windows XP. The application was developed with MATLAB completed by the time-frequency toolbox *tftb* [33].

The approach is composed of two main steps: multi-oriented area detection and text line extraction. Our results are measured according to theses algorithms.

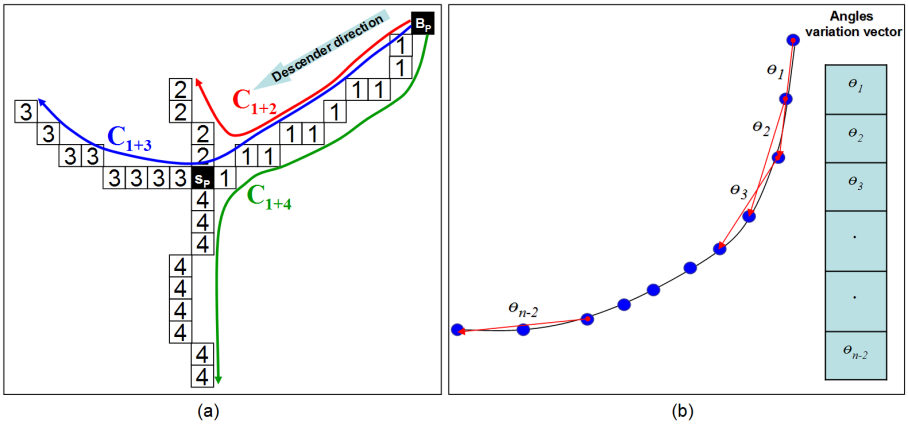


**Fig. 8:** (a) Example of Arabic connected components (the Arabic letter "ra" is connected with the letter "alif"), (b) estimation algorithm of the angular variation.

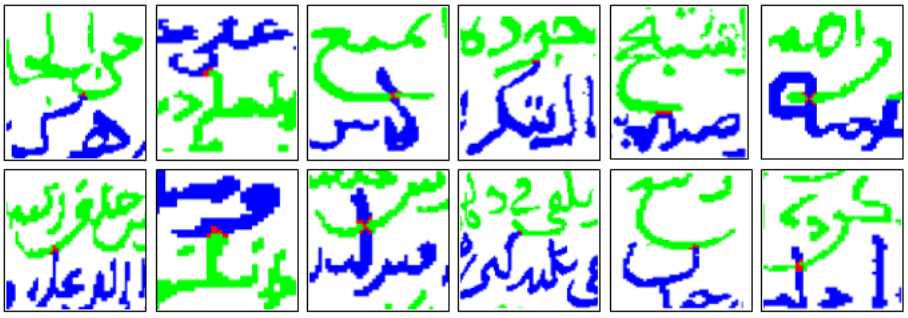

**Fig. 9:** Results of some connected lines separation.

The multi-oriented algorithm is composed of three main steps: image meshing, orientation estimation and orientation extension and paving correction. A global accuracy rate of 97% is reached. The 3% error are shared by the three stages of treatment: 1% is due to the image paving, 1.3% is due to the orientation estimation and 0.7% is due to the orientation extension and paving correction.

In image meshing, it is just needed at least three text lines to obtain a projection profile representing the orientation in each mesh. So, if this criterion is not obtained by the paving algorithm, some errors may happen for area detection. The error rate of 1% is divided in two cases: 0.7% is due to the adjacent line connection and 0.3% is due to the small oriented areas. In the first case, the connection between lines is very frequent in ancient Arabic documents. When the active contour model (Snake) is applied in a mesh to alignements extract, it is possible that it connect two adjacent lines. This will increase the alignement height and consequently the mesh

height. A large mesh may include different oriented areas. In the second case, the oriented areas are composed of few small lines. These areas can be gathered by the paving in other meshes and naturally will not be extracted. The Figure 10 shows the image meshing results of 4 different documents. We can note in these documents the presence at least of three lines in each mesh.

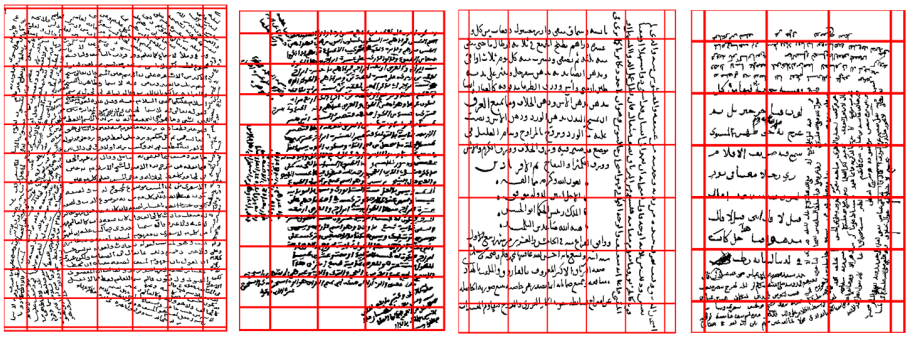


**Fig. 10:** Meshing results of 4 different documents

.

The meshes in our documents have, in some cases, more than one orientation or cursive lines. In the two cases, the orientation estimation is wrong and will be wrong for the orientation extension and consequently for the area detection. The error rate of 0.9% is due to the meshes with multi-orientation. The 0.4% error is due to the Arabic curvature lines in Arabic ancient documents. The Figure 11 shows the results of the first orientation estimation of four documents selected in our database. Each color represent an orientation (see Figure 4 for the color legend). We remark in these documents the presence of meshes with erronous orientation (multi-orientations or cursive lines (Gray color)).

Four extension rules are applied for mesh extension having the same orientation. In the extension phase, any error is happened because all possible orientations in the documents are considered. The error rate of 0.7% is due to the paving correction. As this paving is rectangular, the correction can be applied just along the horizontal and vertical directions. In some cases (oblique areas), the paving correction can not be applied that will yield some segmentation errors. The Figure 12 shows the results of the multi-oriented areas extraction of the four selected documents. Each area is visualized by a color. In theses documents, all the multi-oriented areas are extracted correctely.

Table 3 summarizes the results of the 4 representative documents chosen arbitrary from the 100 documents selected. These results show the effectiveness and the performance of the multi-oriented area detection algorithm.

For line segmentation, the extraction rate reaches 98.6%. The 0.9% of non detected lines is due to the detection area algorithm. The error rate of 0.5% is due to the

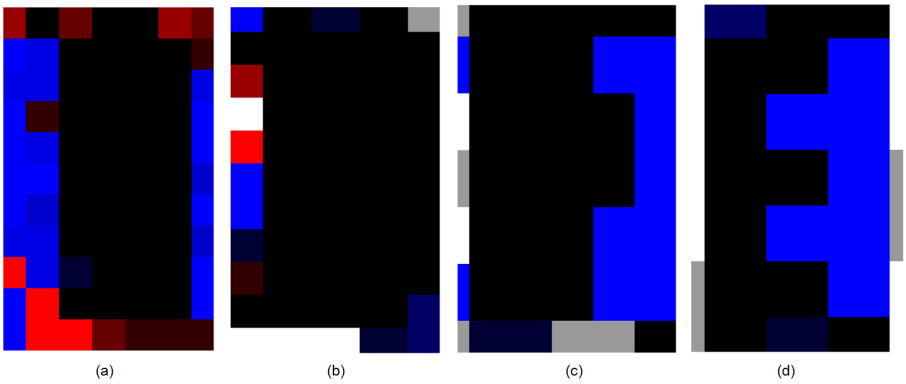


**Fig. 11:** Results of the first orientation estimation of the four selected documents

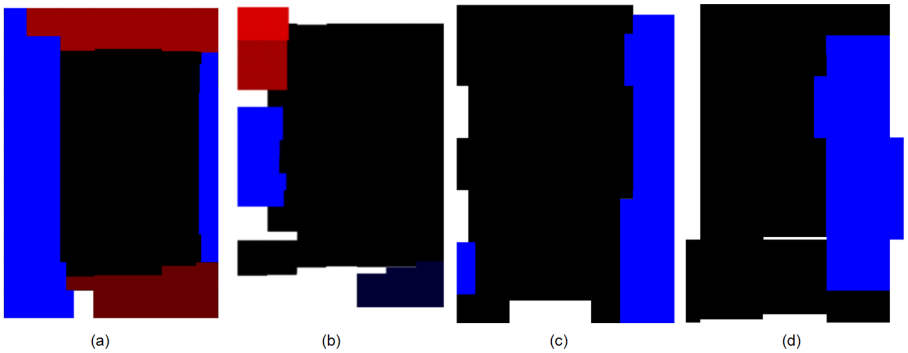


**Fig. 12:** Results of the multi-oriented areas of the four selected documents

.

| Figure | Document size | Resolution (dpi) | $w \times h$ of paving(pixels) | Execution time | Zone number True | Detected |
|---|---|---|---|---|---|---|
| First document | 572×800 | 72 | 75× 75 | 35 s | 5 | 5 |
| Second document | 410×625 | 72 | 75× 75 | 30 s | 5 | 5 |
| Third document | 750×941 | 72 | 120×120 | 34 s | 2 | 2 |
| Fourth document | 362×500 | 72 | 90× 90 | 30 s | 2 | 2 |

**Table 3:** Results of the multi-skew estimation for the four documents.

presence of diacritical symbols in the beginning of lines that create false maxima. Figure 13 illustrates the effectiveness of our algorithm on a sample of 3 documents chosen randomly among the 100 documents processed. To identify the lines, each pair of consecutive lines are presented in two different colors.

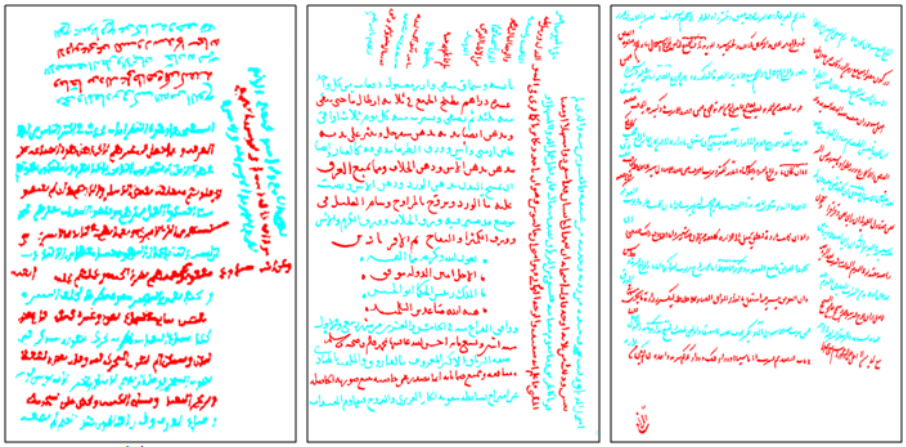


**Fig. 13:** Result examples of the text lines extraction.

## 5 Conclusion

A multi-oriented text line extraction approach is proposed in this chapter based on the local orientation estimation. To extract the lines, the approach proceeds first by an image paving of the document. Then, the orientation in each mesh is estimated, extended and corrected. Finally, the text lines are extracted and separated.

The mesh size is estimated using the active contour model (Snake) approach. This size is fixed once three lines in the mesh are extracted. The skew detection approach use the Cohen's class distributions applied on the projection histogram profile in each mesh and considered as a signal. The Wigner-Ville distribution (WVD) from this class is retained for our application thanks to its interesting properties adapted to the properties of ours signals. The mesh area is extended to similar oriented meshes to obtain largest orientation areas using 4 rules. These rules depends on the orientations presented in these documents. The text lines are extracted in each mesh using a follow-up connected components algorithm. The lines are separated based on the analysis of the terminal Arabic letters.

Experimental results on various types of handwritten Arabic documents show that the proposed method has achieved a promising performance for the text line extraction. This approach will be generalized to other documents types (Latin, Urdu, Farsi etc.) and to heterogeneous documents with text and images.

## References


1. A. Bennasri and A. Zahour and B. Taconet, Extraction des lignes d'un texte manuscrit Arabe, Vision Interface'99, 1999, pp. 42-48.
2. S. Nicolas and T. Paquet and L. Heutte, Text Line Segmentation in Handwritten Document Using a Production System, International Workshop on Frontiers in Handwriting Recognition, 2004, pp. 245-250.
3. A. Antonacopoulos and D. Karatzas, Document Image Analysis for World War II Personal Records, International Workshop on Document Image Analysis for Libraries, 2004, pp. 336-343.
4. V. Shapiro and G. Gluchev and V. Sgurev, Handwritten document image segmentation and analysis, Pattern Recogn. Lett., vol. 14, n. 1, 1993, pp. 71-78.
5. B. Coüasnon, J. Camillerapp, DMOS, une méthode générique de reconnaissance de documents : évaluation sur 60 000 formulaires du XIXe siècle, in Actes du Colloque International Francophone sur l'Écrit et le Document (CIFED'02), Hammamet, 2002.
6. Y. Zheng, H. Li and D. Doermann, A Model-based Line Detection Algorithm in Documents, Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR), 2003, pp.
7. S. Nicolas, T. Paquet and L. Heutte, Text Line Segmentation in Handwritten Document Using a Production System, Proceedings of the 9th Int'l Workshop on Frontiers in Handwriting Recognition (IWFHR-9), 2004, pp.
8. A. Zahour and B. Taconet and S. Ramdane, Contribution à la segmentation de textes manuscrits anciens, Conférence Internationale Francophone sur l'Ecrit et le Document, CIFED'04, 2004, pp.
9. L. Likforman-Sulem and C. Faure, Extracting lines on handwritten documents by perceptual grouping, in Advances in Handwiting and drawing: multidisciplinary approach, C. Faure, P. Keuss, G. Lorette, A. Winter (Eds), 1994, pp.21-38.
10. M. Feldbach and K. D. Tönnies, Line Detection and Segmentation in Historical Church Registers, ICDAR '01: Proceedings of the Sixth International Conference on Document Analysis and Recognition, 2001, pp. 743-748.
11. L. Likforman-Sulem and A. Hanimyan and C. Faure", A Hough Based Algorithm for Extracting Text Lines in Handwritten Document, Proc. of ICDAR'95", 1995, pp. 774-777.
12. Y. Pu, Z. Shi, A natural learning algorithm based on Hough transform for text lines extraction in handwritten documents. In: Proceedings of the 6th International Workshop on Frontiers in Handwriting Recognition, Taejon,Korea, pp. 637-646, 1998.
13. G. Louloudis and B. Gatos and I. Pratikakis and C. Halatsis, Text line and word segmentation of handwritten documents, Pattern Recognition, vol. 42, n. 12, pp. 3169-3183, 2009.
14. Z. Shi and V. Govindaraju, Line Separation for Complex Document Images Using Fuzzy Run length, Int. Workshop on Document Image Analysis for Libraries, 2004.
15. F. LeBourgeois, H. Emptoz, E. Trinh, J. Duong, Networking digital document images. In: 6th International Conference on Document Analysis and Recognition, Seattle, 2001.
16. E. Oztop and A. Y. Mulayim and V. Atalay and F. Yarman Vural, Repulsive attractive network for baseline extraction on document images, Signal Processing, vol. 75, 1999, pp. 1-10.
17. F. Yin, C.-L. Liu, Handwritten text line segmentation by clustering with distance metric learning, Proc. 11th ICFHR, 2008, pp. 229-234.
18. Dod, J.: Effective substances. In: The Dictionary of Substances and Their Effects. Royal Society of Chemistry (1999) Available via DIALOG.
http://www.rsc.org/dose/title of subordinate document. Cited 15 Jan 1999
19. M. Kass and A. Witkin and D. Terzopoulos, Snakes: Active contour models, Proc. 1st ICCV, 987, pp. 259-268.
20. V. Caselles and R. Kimmel and G. Sapiro, Geodesic active contours, International Conference on Computer Vision, 1995, pp. 694-699.
21. C. Pluempitiwiriyawej and J.M.F. Moura and Y. J. L. Wu and C. Ho, STACS: new active contour scheme for cardiac MR image segmentation, IEEE Transactions on Medical Imaging, 2005, vol. 24, n.5, pp.593-603.

22. S. Osher and N. Paragios, Geometric Level Set Methods in Imaging,Vision,and Graphics, 2003, Springer-Verlag New York, Inc.
23. J. A. Sethian, Curvature and the evolution of fronts, Communications in Mathematical Physics, 1985, vol. 101, n. 4, pp. 487-499.
24. F. Leitner and P. Cinquin, From Splines and Snakes to SNAKE SPLINES, Selected Papers from the Workshop on Geometric Reasoning for Perception and Action, 1993, pp. 264-281, Springer-Verlag.
25. R. Ramlau and W. Ring, A Mumford-Shah level-set approach for the inversion and segmentation of X-ray tomography data, J. Comput. Phys., vol. 221, n. 2, pp. 539-557, 2007.
26. S. S. Bukhari and F. Shafait and T. M. Breuel, Segmentation of Curled Textlines using Active Contours, The Eighth IAPR Workshop on Document Analysis Systems (DAS 2008), pp. 270-277.
27. Y. Li and Y. Zheng and D. Doermann and S. Jaeger, Script-Independent Text Line Segmentation in Freestyle Handwritten Documents, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, n. 8, 2008, pp. 1313-1329.
28. D. Mumford and J. Shah, Optimal approximation by piecewise smooth functional and associated variational problems, Commun. Pure Appl. Math., 1989, vol. 42, pp. 577-685.
29. X. Du and W. Pan and T. D. Bui, Text line segmentation in handwritten documents using Mumford-Shah model, Pattern Recogn., vol. 42, n. 12, 2009, pp. 3136-3145.
30. C. Xu and J. L. Prince, Gradient Vector Flow: A New External Force for Snakes, Proc. IEEE Conf. on Comp. Vis. Patt. Recog. (CVPR), 1997, pp. 66-71.
31. J. Montagnat and H. Delingette and N. Ayache, A review of deformable surfaces: topology, geometry and deformation, Journal of Image and Vision Computing, vol. 19, n. 14, 2001, pp. 1023-1040.
32. F. Kaiser James and W. Schafer Ronald, On the Use of the Io-Sinh Window for Spectrum Analysis, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-28, n. 1, 1980.
33. F. Auger and C. Doncarli, Quelques commentaires sur des représentations temps-fréquence proposées récemment", Traitement du Signal, n. 1, vol. 9, pp. 3-25, 1992.
34. L. Cohen, Generalized phase-space distribution functions, J. Math. Phys. n. 5, vol. 7, 1966, pp. 781-786
35. B. Escudié and J. Gréa, Sur une formulation générale de la représentation en temps et en fréquence dans l'analyse des signaux d'énergie finie, CR. Acad. Sci. Paris, 1976, vol. 283, pp. 1049-1051.
36. T.A.C.M. Classen and W.F.G. Mecklenbrauker, The Wigner distribution - A tool for time frequency analysis, Parts I-III, Philips J. Res., 1980, vol. 35, Part I: n°3, p. 217-250; Part II n°4/5, p. 372-389; Part III: n°6, pp. 372-389.
37. F. Hlawatsch and G. F. Boudreaux-Bartels, Linear and quadratic time-frequency signal representation, IEEE Signal Process. Mag, n. 2, vol. 9, pp. 21-67, 1992.
38. P. Flandrin, Time-Frequency/Time-Scale Analysis, Academic Press, San Diego, CA, 1999.
39. E. P. Wigner, On the quantum correction for thermodynamic equilibrium, Phys. Rev., 1932, vol. 40, pp. 749-759.
40. P. Flandrin and W. Martin, A general class of estimators for the Wigner-Ville spectrum of nonstationary processes, Systems Analysis and Optimization of Systems, Lecture Notes in Control and Information Sciences, A. Bensoussan and J-L. Lions (Ed.), pp. 15-23, Springer, Berlin, vol. 62, 1984.
41. R. Ramlau and W. Ring, A Mumford-Shah level-set approach for the inversion and segmentation of X-ray tomography data, J. Comput. Phys., vol. 221, n. 2, 2007, pp. 539-557.
42. E. Kavallieratou, N. Fakotakis, and G. Kokkinakis. Skew angle estimation in document processing using Cohen's class distributions. Pattern Recogn. Lett., 20 :11-13, 1999.
43. E. Kavallieratou, N. Fakotakis, and G. Kokkinakis. Skew angle estimation for printed and handwritten documents using the Wigner-Ville distribution. Image and Vision Computing, 20 :813-824, 2002.

44. N. Ouwayed, A. Belaid and Auger F., Estimation de l'inclinaison d'un document arabe manuscrit numérisé par analyse temps-fréquence des histogrammes de projection, Traitement du Signal, vol. 26, n° 4, p. 0-0, 2009.
45. N. Ouwayed, A. Belaid, "Multi-Oriented Text Line Extraction from Handwritten Arabic Documents", International Workshop on Document Analysis Systems, IAPR, Nara, Japan, 2008.
46. N. Ouwayed, A. Belaid. Separation of Overlapping and Touching Lines within Handwritten Arabic Documents., in "13th International Conference on Computer Analysis of Images and Patterns (CAIP'2009)", IEEE, 2009, p. 237-244.
47. N. Ouwayed, A. Belaid. Une Approache Générale pour l'Extraction des Lignes des Documents Arabes Anciens Multi-orientées, in 12e Colloque International sur le Document Electronique (CIDE.12), 2009.
48. N. Ouwayed, A. Belaid and F. Auger, General Text Line Extraction Approach based on Locally Orientation Estimation, Document Recognition and Retrieval XVII, San Jose, California, 2010.
49. N. Ouwayed, Segmentation en lignes de documents anciens : application aux documents arabes, ,Thèse de doctorat, Université Nancy 2, 2010.